EJERCICIO III.G

```
> Dataset <-
+   read.table("C:/Users/juanjo/Documents/JUAN COMPLETO/Juan acad y otros/UPM/Edificación dep. de matematica aplicada/DOCENCIA edificacion/DOCENCIA Informatica-Montegancedo/econometria ADE/Ejercicios ordenador/salarios_ejercicio_III.G/Wages.csv",
+    header=TRUE, sep=",", na.strings="NA", dec=".", strip.white=TRUE)
```

Apartado 1)

```
> RegModel.1 <- lm(lwage~ed+exp+wks, data=Dataset)

> summary(RegModel.1)

Call:
lm(formula = lwage ~ ed + exp + wks, data = Dataset)

Residuals:
    Min      1Q  Median      3Q     Max
-2.17762 -0.24556  0.00879  0.26802  2.01027

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 5.1272682  0.0667767  76.782  < 2e-16 ***
ed          0.0765796  0.0022746  33.667  < 2e-16 ***
exp         0.0131608  0.0005786  22.746  < 2e-16 ***
wks         0.0064961  0.0012073   5.381 7.83e-08 ***
```

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3993 on 4161 degrees of freedom

Multiple R-squared: 0.2519, Adjusted R-squared: 0.2514

F-statistic: 467.1 on 3 and 4161 DF, p-value: < 2.2e-16

El plano de regresión es

lwage=5.13+0.076ed+0.013exp+0.006wks

Las tres variables son independientemente muy significativas para los salarios, con p-valores prácticamente nulos la educación y la experiencia, y muy pequeño la antiguedad. También conjuntamente son las tres muy significativas: p-valor del test F es también nulo. Se observa que de las tres variables la que influye más en el salario son los años de educación recibida, seguida de la experiencia y finalmente la antiguedad en el puesto de trabajo.

Apartado 2)

Definimos nueva variable sex_n para pasar el género a variable numérica:

```
> Dataset$sex_n <- with(Dataset, as.numeric(sex))
```

Observamos que esta variable tiene dos valores 1 (female) y 2 (male). Por tanto, el coeficiente (pendiente) del plano de regresión (valor estimado del parámetro beta_sex_n correspondiente) si es positivo indica discriminación de género en contra de la mujer, sólo por el hecho de serlo (céteris páribus) (¿qué indicaría si fuese negativo?).

```
> RegModel.2 <- lm(lwage~ed+exp+sex_n+wks, data=Dataset)

> summary(RegModel.2)

Call:
lm(formula = lwage ~ ed + exp + sex_n + wks, data = Dataset)

Residuals:
    Min      1Q  Median      3Q     Max
-2.19860 -0.23595  0.00272  0.25088  1.94862

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 4.4632175  0.0691061  64.585  < 2e-16 ***
ed          0.0754246  0.0021419  35.214  < 2e-16 ***
exp         0.0119175  0.0005473  21.774  < 2e-16 ***
sex_n       0.4298585  0.0185805  23.135  < 2e-16 ***
wks         0.0041945  0.0011409   3.677 0.000239 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3759 on 4160 degrees of freedom
Multiple R-squared:  0.3372,     Adjusted R-squared:  0.3366
```

F-statistic: 529.1 on 4 and 4160 DF, p-value: < 2.2e-16

Vemos que por el test t las cuatro variables son independientemente significativas, siendo las tres primeras muy significativas (p-valores prácticamente nulos) y la antiguedad notablemente significativa. También las cuatro variables son conjuntamente muy significativas. Es llamativa la discriminación de género en contra de la mujer, ¡muy superior respecto al resto de variables explicativas!

Apartado 3)

Estudiamos ahora el que llamamos modelo completo, añadiendo la variable raza (negro o no).

```
> Dataset$black_n <- with(Dataset, as.numeric(black))
```

Observamos que la asignación numérica es: 1 (no es negro), 2 (es negro). Por tanto, si hay discriminación salarial en contra de los trabajadores de raza negra, respecto al resto, el coeficiente estimado deberá ser negativo.

```
> RegModel.3 <- lm(lwage~black_n+ed+exp+sex_n+wks, data=Dataset)

> summary(RegModel.3)

Call:
lm(formula = lwage ~ black_n + ed + exp + sex_n + wks, data = Dataset)

Residuals:
     Min       1Q   Median       3Q      Max
-2.20174 -0.23548  0.00406  0.25088  1.94269
```

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  4.7082229  0.0774416  60.797  < 2e-16 ***
black_n     -0.1583203  0.0230638  -6.864 7.66e-12 ***
ed           0.0737809  0.0021435  34.421  < 2e-16 ***
exp          0.0120483  0.0005447  22.121  < 2e-16 ***
sex_n        0.4025590  0.0189015  21.298  < 2e-16 ***
wks          0.0040834  0.0011347   3.599 0.000324 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3738 on 4159 degrees of freedom
Multiple R-squared:  0.3446,     Adjusted R-squared:  0.3438
F-statistic: 437.4 on 5 and 4159 DF,  p-value: < 2.2e-16
```

Las otras variables se mantienen con t-valores y significaciones (p-valores) similares a las obtenidas en el apartado 2). También las cinco variables son conjuntamente muy significativas (p-valor del F-test del modelo constante es <1.2*10^(-16): claramente se rechazan todas las hipótesis nulas de que los cinco coeficientes sean nulos tanto conjuntamente (F-test), como individualmente (t-tests). Como era de esperar, se observa una discriminación en contra de los trabadores de raza negra respecto a otras razas (céteris páribus: coeficiente -0.16), siendo además la segunda causa de discriminación por detrás del sexismo (coeficiente 0.4, que prácticamente se mantiene respecto al modelo del apartado anterior).

Apartado 4) Comparación de los modelos de los apartados 1) y 3):

Modelos->Test de hipótesis-Comparar dos modelos->marcar los modelos , en nuestro caso son RegModel.1 (modelo restringido) y RegModel.3 (completo)

(así los habímos denominado al obtenerlos).

```
> anova(RegModel.1, RegModel.3)

Analysis of Variance Table

Model 1: lwage ~ ed + exp + wks

Model 2: lwage ~ black_n + ed + exp + sex_n + wks

  Res.Df    RSS     Df    Sum of Sq     F        Pr(>F)

1   4161   663.48

2   4159   581.26    2     82.218    294.14    < 2.2e-16 ***

---

Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Mediante el test F, con un F-valor de 294.14, para F_{g,n-k}=F_{2,4159} (g=2, número de variables que quito, n=4165, k=6), con un p-valor prácticamente nulo, rechazamos claramente la hipótesis

H0: beta_{sex_n}=beta_{black_n}=0,

es decir, que la discriminación por género y por raza negra no sea conjuntamente significativa, para explicar las variaciones en los salarios. En consecuencia, el modelo más significativo es el completo con las cinco variables explicativas. De hecho esto ya se podía intuir viendo la significación individual tan clara de esas dos variables discriminatorias en el modelo completo. De obligarnos a quitar alguna variable, sería en todo caso la de la antiguedad, que es la que da un test t con p-valor mayor,pero todavía muy lejos del 0.05...

Apartado 5)

Observamos que con los coeficientes estimados en el modelo completo de las variables sex_n y edu (que es la siguiente variable en afectar a los salarios):

0.4025590/0.0737809=5.45.

Luego parece razonable plantearse la hipótesis de que la que haya una relación \beta_s=5.5\beta_e (escribiendo s por sex_n y e por ed)

H0:\beta_s=5.5\beta_e

Esto daría lugar a un modelo restringido, sustituyendo esas dos variables por una nueva variable z=5.5*sex_n+edu, y obteniendo la regresión del restringido (estadisticos->ajuste de modelos..):

```
> Dataset$z <- with(Dataset, 5.5* sex_n+ ed)


> RegModel.4 <- lm(lwage~black_n+exp+wks+z, data=Dataset)


> summary(RegModel.4)


Call:

lm(formula = lwage ~ black_n + exp + wks + z, data = Dataset)


Residuals:

    Min      1Q  Median      3Q     Max

-2.20104 -0.23555  0.00423  0.25035  1.94227


Coefficients:

            Estimate Std. Error t value Pr(>|t|)

(Intercept)  4.7064390  0.0764536  61.559  < 2e-16 ***

black_n     -0.1579199  0.0228960  -6.897  6.1e-12 ***

exp          0.0120323  0.0005335  22.555  < 2e-16 ***
```

wks 0.0040703 0.0011310 3.599 0.000323 ***

z 0.0736162 0.0018192 40.466 < 2e-16 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3738 on 4160 degrees of freedom

Multiple R-squared: 0.3446, Adjusted R-squared: 0.344

F-statistic: 546.9 on 4 and 4160 DF, p-value: < 2.2e-16

Observación: no lo piden, pero como $R^2=0.3446=R_R^2$ (coeficientes del completo y de este restringido), de aquí intuimos que el F-valor asociado a H0 (ojo: no es el F-valor de la tabla anterior, que es el F-valor del test global de las variables del modelo anterior) va a salir muy pequeño , de hecho 0 en la aproximación aquí,

F= (n-k)/g (R^2-R_R^2)/(1-R^2)=0 , n-k=4159, g=1 (una restricción).

Por tanto, ya anticipamos que no podremos rechazar H0.

Ahora para hacer el F-test de la hipótesis H0, hemos de comparar con el completo:

> anova(RegModel.4, RegModel.3)

Analysis of Variance Table

Model 1: lwage ~ black_n + exp + wks + z

Model 2: lwage ~ black_n + ed + exp + sex_n + wks

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	4160	581.26				
2	4159	581.26	1	0.0029516	0.0211	0.8845

Luego claramente no se pude rechazar (con un p-valor tan grande, *se dice a veces que se ha demostrado estadísticamente H0...) la hipótesis*

H0: la influencia explicativa de la sobre los salarios (lwages) de la discriminación por género es del orden de un factor del orden de 5.5 veces superior a la de la educación.